An Initial Comparison of Methods for Representing and Aggregating Experimental Uncertainties involving Sparse Data[1]

Vicente Romero[2], Laura Swiler, Angel Urbina
Sandia National Laboratories,[3] *Albuquerque, NM*

Abstract

This paper discusses the handling and treatment of uncertainties corresponding to relatively few data samples in experimental characterization of random quantities. The importance of this topic extends beyond experimental uncertainty to situations where the derived experimental information is used for model validation or calibration. With very sparse data it is not practical to have a goal of accurately estimating the underlying variability distribution (probability density function, PDF). Rather, a pragmatic goal is that the uncertainty representation should be conservative so as to bound a desired percentage of the actual PDF, say 95% included probability, with reasonable reliability. A second, opposing objective is that the representation not be overly conservative; that it minimally over-estimate the random-variable range corresponding to the desired percentage of the actual PDF. The performance of a variety of uncertainty representation techniques is tested and characterized in this paper according to these two opposing objectives. An initial set of test problems and results is presented here from a larger study currently underway.

I. Introduction

This paper discusses and tests various statistical concepts and techniques for expressing uncertainty due to random variability (aleatory uncertainty) when limited data samples exist of the random quantity of interest. Limited sampling introduces an epistemic contribution of uncertainty to the problem of random-variable characterization. The importance of this topic extends beyond experimental uncertainty characterization to situations where the derived experimental information is used for model validation or calibration purposes.

If the random quantity is fully sampled such that its characteristic probability density function (PDF) is fully known, then well-known approaches can be used to express and work with the probabilistic uncertainty. These include Monte Carlo and Quasi Monte Carlo approaches (often involving constructed response surfaces for non-linear response functions, e.g. references [1]-[4]); linearized-response function approaches popular in experimental uncertainty quantification (e.g. [5]-[10]); and more recent approaches like polynomial chaos and stochastic expansion methods for uncertainty propagation and aggregation (e.g. [11]).

However, when relatively few samples are available, a substantial epistemic contribution of uncertainty exists in addition to the aleatory uncertainty due to the quantity's random variability. This epistemic uncertainty undermines accurate estimation of the underlying variability distribution (PDF). Hence, it is not practical to pursue a goal of accurate PDF estimation from sparse data. Rather, a pragmatic goal is that the uncertainty representation should be ***conservative*** so as to bound a desired percentage of the actual PDF, say 95% included probability, with reasonable reliability. A second, opposing objective is that the representation not be overly conservative; that it minimally over-estimate the random-variable range

---

[1] This paper is a work of the United States Government and is not subject to copyright protection in the U.S.
[2] AIAA Senior Member, corresponding author: vjromer@sandia.gov
[3] Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

corresponding to the desired percentage of the actual PDF. The presence of the two opposing objectives makes the sparse-data uncertainty representation problem an interesting and difficult one.

A classical approach to working with sparse random data is to use statistical tolerance intervals [12] to provide a reliably conservative interval representation for the combined epistemic and aleatory uncertainty associated with the limited data. This approach is tested in this paper, along with a more recent kernel-density estimation technique [13] for constructing conservative PDF representations from sparse data. Another non-parametric approach [14] is also investigated, along with the very common practice of simply fitting the sample data with a normal distribution. Taking a cue from [15] and [16], an approach was also tried that uses a Johnson-family four-parameter representation of PDFs ([17]). However, optimization difficulties related to estimating the parameters of the PDF could not be overcome in time for this paper. It is anticipated that these difficulties will be surmounted in the near future, for follow-on studies that will complement the material in this paper.

The four approaches pursued in this paper are described more fully in section III, following a description in section II of the uncertainty-representation problem and the test plan for assessment of the approaches. Section IV presents results and performance comparisons of the four methods. Section V provides some discussion and conclusions of the present study. It is important to note that this paper reports only preliminary methodology and results from a work in progress. A more refined study is currently underway.

II. Sparse Data Uncertainty Representation Problem Description and Test Plan

The larger problem for study is shown in Figure 1. Model PDFs of normal, uniform, and right-triangular diverse shapes are to be randomly sampled and the samples are to be used by the methods under investigation to identify a range that ostensibly covers the 0.025 to 0.975 percentile ranges of the model PDFs. The PDFs are sized and located relative to the r_{e_o} reference lines shown in the figure such that the range between the said percentiles is the same length for all PDFs and the range is centered on the reference lines. In the present study r_{e_o} (the reference line) is set to a value of zero and the 0.025 and 0.975 percentile range is set to 2. Accordingly, the 0.025 and 0.975 percentiles of the normal, uniform, and triangular PDFs reside at respective values of -1 and 1.

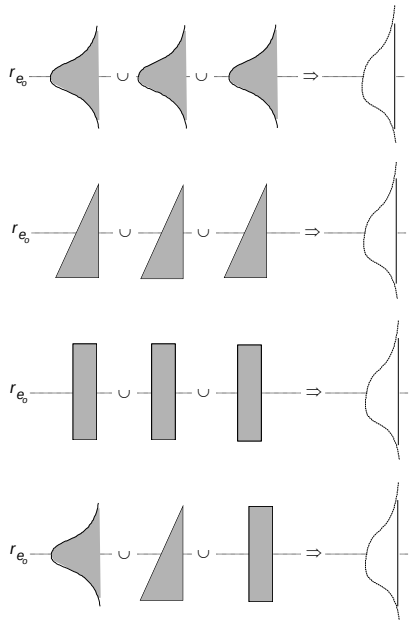

Figure 1. **Test matrix for larger study of PDF representation performance of cited methods, under limited sample data from normal, uniform, and triangular test PDFs as shown. Initial study in this paper involves only top line of normal PDFs in this figure.**

The investigated methods are tested for their performance in estimating, from random samples of the PDFs, ranges that contain the true 0.025 to 0.975 percentile range (-1,1). The estimated ranges are compared to the true ranges and method performance is assessed and characterized according to the metrics and procedures presented in section IV of this paper.

It was initially envisioned that the study would be conducted for 2 samples from each PDF, then a doubling of samples to 4 per PDF, then 8, 16, 32, and finally 64 samples of each PDF. By doubling sample size each time, a sense of the rates of improvement (convergence rates) of the various small-sample uncertainty representation schemes could perhaps be obtained. However, given time constraints, the present study involves sampling sizes of n = 2, 8, and 32. Since the increase in sample size is a constant factor (of 4) each time, a rate of decrease of method estimation errors with increasing sample size can be calculated. This was not done for the preliminary study reported in this paper.

The study also assesses method performance in the context of fitting multiple sources of aleatory uncertainty (from samples of multiple PDFs) that aggregate to a total resultant uncertainty for a system. It was surmised that the aggregation of multiple sources of uncertainty may smooth out and reduce, or alternatively could amplify, the performance differences between the methods—so the test plan investigates this aspect. Since it is often the case that only a handful of empirical samples of each source distribution exists, will the sparse-sample methods avoid a severe underestimation of the combined variability in the resultant exact PDFs at right in the figure? In the context of this study, do the methods avoid underestimation of the true 0.025 to 0.975 percentile range of the combined variability at the system level?

Experience suggests that typically any given system response quantity of project-level engineering interest is dominated by a few random variables even if many stochastic variables exist in the system. Therefore the study was designed to consider three equally dominant random variables having equal impact on the variability of system response. A system response that is linear in the ranges of the uncertain variables is considered in the study in order to avoid any confounding of sparse-sample methodology results with errors from any inexactly captured non-linearities in uncertainty propagation procedures that would have to be employed. Uncertainty propagation procedures are avoided altogether by specifying that the three random variable sources are independent, and that system response is linear in the ranges over which the random variables vary. The problem then defaults to a linear convolution problem as depicted in Figure 1. As shown, four different convolution problems are planned for study, each with a different combination of three source uncertainties of normal, uniform, and/or triangular PDF shapes.

The initial study reported in this paper involves only the top line in the figure (only normal PDFs). The normal PDFs in the figure have a mean μ = zero and a standard deviation $\sigma = 1/1.96$ such that the $\mu \pm 1.96\sigma$ extents of the normal PDFs, which mark their 0.025 to 0.975 percentiles, have values of -1 and 1 as previously stipulated. A convolution of the three normal PDFs in the top line of the figure yields a resultant normal PDF with zero mean and a variance σ_{res}^2 that is the sum of the three contributing variances σ^2. The resultant standard deviation is $\sigma_{res} = [(1/1.96)^2 + (1/1.96)^2 + (1/1.96)^2]^{1/2}$. The 0.025 to 0.975 percentiles of the resultant normal PDF therefore lie at $0 \pm 1.96\sigma_{res} = (-1.732, 1.732)$.

The following procedure was undertaken for the present study.

1. Randomly sample the normal PDF for a specific number of samples prescribed in the test plan (n = 2, 8, or 32). Do this three times, once for each appearance of the normal PDF in the top line of Figure 1.
2. For one of the four methods being evaluated, estimate from the three sets of n samples (obtained in step 1) three ranges that ostensibly bound the 0.025 to 0.975 percentile range of the exact PDF from which the samples are drawn.
3. Use the three estimated PDF percentile ranges, one for each normal PDF in Figure 1, to estimate the aggregate 0.025 to 0.975 percentile range of the PDF resulting from convolving the three normal PDFs in the figure. The aggregate percentile range estimation procedure for each method is explained in section III.
4. Using metrics and procedures from Section IV, assess and characterize the estimated percentile range against the true 0.025 to 0.975 percentile range (-1.732, 1.732) of the exact resultant PDF.

5. Perform steps 1 – 4 1000 times (1000 "*trials*") to characterize the random variability of method performance under different sets of random samples of the three normal PDFs.
6. Upon completion of step 5, 3000 estimated bounds will exist for the 0.025 to 0.975 percentile range of the source PDFs, Normal($\mu = 0$, $\sigma = 1/1.96$). Assess these 3000 values against the true range (-1,1) using the metrics and procedures presented in Section IV.
7. Perform steps 1 - 6 for each sparse-data treatment approach.
8. Perform steps 1 – 7 for $n = 2$, 8, and 32 samples per source PDF.

III. Uncertainty Estimation Method Descriptions

A. Percentile Estimation by Fitting Normal PDF to the Sample Data

A very common approach in engineering practice is to simply fit the sample data with a normal PDF. This practice will be analyzed in Section IV for effectiveness with respect to the performance objectives previously discussed and quantified in Section IV. For the present purposes the 0.025 and 0.975 percentiles of the fitted normal PDF are used to estimate the true 0.025 and 0.975 percentile range (-1,1) of the sampled PDF.

To provide an estimated bounding range for the exact 0.025 to 0.975 percentiles of the resultant PDF from convolving the three source PDFs depicted in the top line of Figure 1, the following procedure is used. For the three normal PDFs produced per "trial" (step 5 in Section II), their means are added to obtain the value of the mean of the resultant PDF of the three convolved trial PDFs. This shortcut makes use of a standard rule for convolution of PDFs. Likewise, the variance of the convolved product PDF is obtained by summing the variances of the contributing source PDFs. Hence, the mean and variance are easily obtained for the PDF resulting from convolving the three trial PDFs. It is also a fact that convolving normal PDFs produces a resultant PDF that is normally distributed. Because it is Normal, the easily obtained mean and variance of the resultant PDF fully define its density function. The 0.025 to 0.975 percentile range of the aggregate PDE so defined in each trial is then compared against the exact 0.025 to 0.975 percentile range (-1.732,1.732) of the true aggregate PDF to yield the error characterization presented in Section IV.

B. Tolerance Interval Method

Another simple approach is to use statistical tolerance intervals [12] to provide a reliably conservative interval representation for the combined epistemic and aleatory uncertainty associated with sparse data. This approach provides multiplication factors to scale the calculated standard deviation from a sparse data set to yield specific "tolerance" intervals advertised to, at a desired user-specified level of confidence, span a particular percentile range of the sampled PDF. For the purposes here a 90% confidence level is specified and the percentile range of coverage is selected to be 95% to target the 0.025 to 0.975 percentile range of the exact PDF. The associated tolerance interval is constructed by multiplying the calculated standard deviation σ by the following factors f in Table 1 to create an interval of total length $2f\sigma$, where the interval is centered about the calculated mean of the data samples. The produced tolerance intervals are compared in Section IV to the true 0.025 and 0.975 percentile range (-1,1) of the sampled PDF.

Table 1. Tolerance Interval Factors versus Number of Samples of PDF

	$n = 2$ samples	$n = 8$ samples	$n = 32$ samples
factor	18.8	3.264	2.395

To provide an estimated bounding range for the exact 0.025 to 0.975 percentiles of the resultant true convolved PDF depicted in the top line of Figure 1, the following procedure is used. For the three tolerance intervals produced per trial (step 5 in Section II), their means are added to obtain the value of a resultant mean. This gives the center of a net tolerance interval that is compared in Section IV to the exact 0.025 to 0.975 percentile range range (-1.732,1.732) of the true aggregate PDF. Using the classical result of Kline & McClintock [5], the length of the net tolerance interval is the root sum of squares of the three tolerance intervals calculated for that particular trial.

C. PDF Estimation by the Pradlwarter-Schuëller Kernel Density Method

Kernel density estimation (KDE, seminal papers by Rosenblatt [18], and Parzen [19]) is a technique used to estimate the density of a random variable X given n independent samples X_1, ..., X_n of it. Let $K(.)$ be a kernel function which is a non-negative, real-valued, symmetric function that integrates to one $\int K(x)dx = 1$. Then the KDE is

$$f^{\{KDE\}}(x) = \hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - X_i}{h}). \qquad (1)$$

Here, h is known as the *bandwidth parameter* which controls the influence of each sample in providing a density estimate at a near-by point. Small h corresponds to a small region of influence; a large h to a large one. Common kernel functions include the Gaussian, uniform, triangular, and Epanechnikov kernels. In this paper, we use a Gaussian kernel.

Selection of the bandwidth parameter is the subject of a vast literature. Popular methods include cross validation (see [20], [21]) and asymptotic analysis (see [20], [22]). A common approach to measuring the error in the estimation process is the mean integrated squared error (MISE):

$$MISE = E\int (\hat{f}_h(x) - f(x))^2 dx. \qquad (2)$$

However, it is not feasible to minimize the MISE with respect to the bandwidth h unless the true density function $f(x)$ is known—which is not usually the case. Thus, in practice, people use cross-validation measures. In this approach, one maximizes a criterion such as the maximum-likelihood cross validation (MLCV) measure. At each observation X_i, the log-likelihood of the density at this point X_i is estimated based on all remaining observations except X_i. These log-likelihoods are then averaged over all observations:

$$MLCV = \frac{1}{n}\sum_{i=1}^{n} \log(\hat{f}_{-i}(X_i)) . \qquad (3)$$

Bandwidth estimation needs to be automated in some fashion. However, the use of automated methods such as optimization methods which maximize a log likelihood function (for example) may be subject to local optima depending on the starting bandwidth. Figure 2 below shows two cases where a kernel density estimate was constructed over the same data set of four points, denoted with red X marks. The kernel density estimate shown with the green line is a KDE constructed with a small bandwidth (bandwidth value of 0.2), resulting in multimodal behavior of the output where the modes are around the actual data points. The density estimate shown with the blue line is constructed using a large bandwidth in the kernel (0.99), resulting in a smoothing of the density estimate.

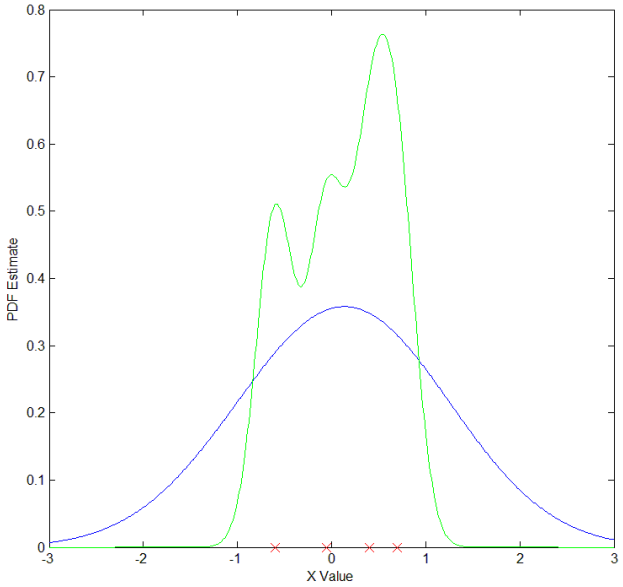

Figure 2. Kernel Density Estimate based on four points at the red X marks. The green line is a KDE constructed with a small bandwidth. The blue line is a KDE constructed with a large bandwidth.

In this work we did not want to use the cross-validation methods because of the small data sets (e.g. "leave one out" when you only have two data points leaves only one data point with which to construct the density estimate). We wanted to have an approach that was robust to small numbers of data points. Also, we wanted to ensure that density estimates based on the sparsest data sets were the most conservative. That is, we wanted to ensure that the tails of distributions derived from fewer data points would encompass or envelop the tails of the distributions derived from more data points. As Pradlwarter & Schuëller ([13]) state "In general, this condition is not fulfilled." To handle low numbers of data points robustly, we used a bandwidth estimation approach that they proposed and implemented in [13]. This approach was designed to minimize the probability that future measurements will lead to data points outside the domain of existing data points. Specifically, Pradlwarter & Schuëller define the bounds *a* and *b*:

$$a = x_{\min} - \frac{1}{2n-2}(x_{\max} - x_{\min})$$
$$b = x_{\max} + \frac{1}{2n-2}(x_{\max} - x_{\min}) \qquad (4)$$

where x_{min} and x_{max} are the minimum and maximum values of the existing *n* data points. Given these bounds and the assumption that the existing data points are equivalent to *n* independent and identically distributed realizations of sample points from an unknown distribution $f(x)$, the probability that a point will fall outside the bounds *a* and *b* is:

$$p = 1 - \int_a^b f(x)dx\,. \qquad (5)$$

The probability that *n* independent data points will fall inside the domain is $\left(\int_a^b f(x)dx\right)^n$. To achieve a confidence level (1-α) that the probability of a new point falling outside the bounds will not exceed *p*, Pradlwarter and Schuëller suggest that interpreting $\alpha = (1-p)^n$ as the confidence level or level of significance, and thus $P(\alpha,n) = 1 - \alpha^{1/n}$. Given this framework, the goal is to find a bandwidth *h* that will satisfy the following condition:

$$\int_{-\infty}^{a} f(x;h)dx + \int_{b}^{+\infty} f(x;h)dx = P(\alpha, n)\,. \qquad (6)$$

They use a Gaussian kernel, so that the KDE is specified as:

$$f^{\{KDE\}}(x) = \hat{f}(x;h) = \frac{1}{nh\sqrt{2\pi}} \sum_{i=1}^{n} \exp\left(\frac{x - X_i}{2h^2}\right)\,. \qquad (7)$$

We found this formulation to be easy to optimize. We specified α=0.1, and then calculated P(α,*n*) for a given number of data points *n*. Then, we found *h* which minimized the following expression:

$$abs\left[\left(\int_{-\infty}^{a} f(x;h)dx + \int_{b}^{+\infty} f(x;h)\right) - P(\alpha, n)\right] \qquad (8)$$

This is an easy optimization calculation, and appears robust to a limited number of data points. The KDE densities constructed in this fashion have the property that the density estimates are narrower for increasing number of points: the density estimates based on a few data points encompass the density estimates based on more points. Also, the density estimates are wide and smooth when constructed with only a few points.

We constructed the KDE percentile range estimates by calculating the KDE values at input values between -5 and 5, at increments of 0.01. We took the KDE PDF and calculated the CDF (cumulative density function) from it. Then we constructed 10,000 random samples and interpolated where they would fall on the CDF curve (that is, we did an inverse mapping from the CDF back to the X values for the 10K samples). These 10K samples of a given KDE PDF formed the basis for locating its 0.025 to 0.975 percentiles. The percentile range was then compared against the exact 0.025 to 0.975 percentile range (-1,1) of the true PDF to yield the error characterization described in Section IV. To provide an estimated bounding range for the exact 0.025 to 0.975 percentiles of the resultant PDF from convolving the three source PDFs depicted in the top line of Figure 1, the 10K samples from each of the three KDE PDFs per "trial" (step 5 in Section II) were added in the following manner. The 1st sample from each of the three KDE PDFs was taken and the values of the three samples were added together. This sum constituted one sample of the aggregate PDF corresponding to a convolution of the three KDE PDFs. Similarly, 9999 other such samples of the resultant PDF were constructed. The resulting 10K samples portray the aggregate PDF for a given trial, 1000 of which trials were ultimately performed in the study. For each trial the 0.025 to 0.975 percentile range of the aggregate KDF PDE was compared against the exact 0.025 to 0.975 percentile range (-1.732,1.732) of the true aggregate PDF to yield the error characterization described in Section IV.

D. PDF Estimation by the Sankararaman-Mahadevan Non-Parametric Method

The following methodology was developed by Sankararaman and Mahadevan [14] for the use of non-parametric distributions to fit point data. The following summary of the technique is paraphrased from [14]. For a more complete explanation of this approach see that reference.

Discretize the domain of *X* into a finite number of points, say $\theta_i, i = 1{:}\,Q$. The domain is chosen based on the available data; the lowest value and the highest value are chosen as the lower bound and the upper bound of the domain, respectively. Assume that the PDF values at each of these *Q* points are denoted by $f_X(x = \theta_i) = p_i \text{ for } i = 1{:}\,Q$. Using an interpolation technique, the entire PDF $f_X(x)$ can be approximated for all $\theta \in X$, i.e. over the entire domain of *X*. Then the probability of observing the given point data is given by the likelihood, ***L(p)***. This likelihood is a function of the following: (a) The discretization points selected, i.e. $\theta_i, i = 1{:}\,Q$. (b) The corresponding PDF values p_i; and (c) The type of interpolation technique used. For this work, the discretization is fixed, i.e. uniformly spaced p_i values $(i = 1{:}\,Q)$ over the domain of *X* are chosen in advance and the likelihood is maximized over

the various values of p_i. The value of *Q* (number of discretization points) is chosen based on computational power—the larger the *Q*, the finer (in terms of flexibility) is the resulting interpolation of the PDF. In this paper, for the purpose of illustration, *Q* has been chosen to be equal to 11. Also, θ_1 is equal to the minimum of the available data (Xmin) and θ_Q is equal to the maximum of the available data (Xmax), and the intermediate $\theta's$ are uniformly interspersed between Xmin and Xmax.

The optimization problem is formulated as:

$$\begin{aligned} &\text{Given } \theta_i \in X \forall i, i = 1{:}\,Q \\ &Max\ L(\boldsymbol{p}) \\ &where\ \boldsymbol{p} = \{p_1, p_2, p_3 \dots p_{Q-1}, p_Q\}\ \&\ f_X(x = \theta_i) = p_i \end{aligned} \tag{9}$$

subject to:

$(1)\ p_i \geq 0 \forall i$

$(2)\ f_X(x) \geq 0 \forall x$

$(3) \int f_X(x)dx = 1$

Note: $p_i\ at\ \theta_i\ (i = 1{:}\,Q)$ is used to interpolate and calculate $f_X(x)$.

This optimization problem maximizes the likelihood function subject to three constraints. The first constraint states that the vector **p** (that contains probability values) needs to be positive. The second and third constraints state that $f_X(x)$ must be positive and the area under this curve should be equal to unity, so as to satisfy the properties of a PDF. The PDF may be constructed using various interpolation methods. Several interpolation techniques, such as linear interpolation, spline-based interpolation and Gaussian process interpolation could be used. In this paper a cubic spline is used for interpolation. The term ‘‘spline’’ refers to a wide variety of functions used for purposes of data interpolation and/or smoothing. Spline functions for interpolation are calculated as the minimizer of some suitable measure (e.g., the integral of the squared curvature) subject to some constraints (the interpolation constraints). For a more detailed explanation of splines, refer to Ahlberg et al. [23]. For the example presented here, eleven discretization points are chosen and the PDF is constructed. Then a similar methodology to that for the Pradlwarter-Schuëller method is followed as described in the last paragraph of section III.C.

IV. Uncertainty Estimation Results and Performance Comparisons

A. Measures of Uncertainty Estimation Error

Figure 3 defines discrepancy or error quantities to be calculated from comparing the exact 0.025 to 0.975 percentile range of the true PDF against the percentile range calculated from the estimation methods being tested. For the *i*th trial the calculated discrepancies at the upper and lower ends of the exact percentile range are given by:

$$\begin{aligned} \Delta_{U\text{-}i} &= r_{U\text{-}i} - r_{U\text{-}exact} \\ \Delta_{L\text{-}i} &= r_{L\text{-}exact} - r_{L\text{-}i}\ . \end{aligned} \tag{10}$$

These equations return a positive value of discrepancy $\Delta_{L\text{-}i}$ when the estimated range extends beyond or "bounds" the true percentile at the lower end. The discrepancy $\Delta_{U\text{-}i}$ at the upper end is similarly positive when the estimated range extends beyond the true percentile at the upper end. In the opposite case where the true range extends beyond the estimated range, the discrepancy Δ is a negative value.

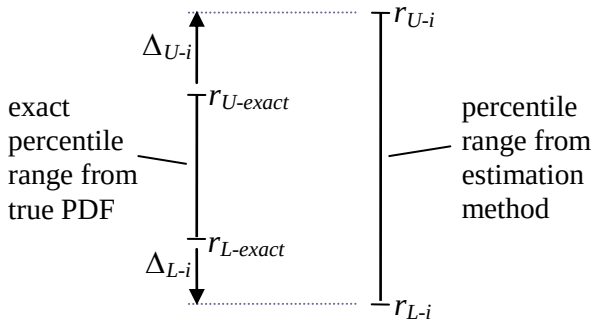

Figure 3. Definition of discrepancy or error between true percentile range and range yielded by estimation method.

A primary goal of sparse-sample methodology is that the calculated percentile range be conservative so as to envelope or bound the actual percentile range of the sampled PDF. Therefore, positive discrepancy values Δ from over-prediction of the actual percentile range are looked upon as more desirable errors than are negative values of Δ. Hence, we are looking for methodologies that will produce positive values of $\Delta_{U\text{-}i}$ and $\Delta_{L\text{-}i}$ with high reliability (i.e., in a high percentage of the trials). Nonetheless, a second opposing objective is that the positive values of Δ be small; that the method minimally over-estimate the true percentile range of the sampled PDF. This presence of the two opposing objectives makes the sparse-data uncertainty representation problem an interesting and difficult one. A clear "winner" would be the estimation method that encompasses the exact result the most times in the trials and simultaneously has the least overshoot error (smallest average and maximum positive Δ values). Unfortunately, it will be seen that no method enjoys the best of both worlds, but some score significantly better than others over the set of attributes or measures of method performance characterized in the following.

B. Presentation of Raw Results—Histograms of Estimation Errors

Figures 4-7 show the raw discrepancy results for the various estimation methods, numbers of data samples taken, and the particular test case: 95 percentile estimation for individual normal PDF (3000 trials), or for the convolved PDF (1000 trials). On each plot the discrepancy results are separated into three histograms with integrated areas proportional to the number of trail results binned into the histograms as follows:

- the green histogram contains the +/+ or "pos/pos" trial sums S_i (= $\Delta_{U\text{-}i}$ + $\Delta_{L\text{-}i}$) where both $\Delta_{U\text{-}i}$ and $\Delta_{L\text{-}i}$ were positive in a trial. These results represent a trial in which the estimated interval bounded the true interval.
- the red histogram contains the -/- or "neg/neg" trial sums S_i (= $\Delta_{U\text{-}i}$ + $\Delta_{L\text{-}i}$) where both $\Delta_{U\text{-}i}$ and $\Delta_{L\text{-}i}$ were negative in a trial. These results represent a trial in which the estimated interval fell short of the the true interval range at both upper and lower ends of the range.
- the blue histogram contains the "mixed" trial sums S_i (= $\Delta_{U\text{-}i}$ + $\Delta_{L\text{-}i}$) where one of $\Delta_{U\text{-}i}$ and $\Delta_{L\text{-}i}$ was positive and the other was negative. These results represent a trial in which the true interval is bounded by the estimated interval at only one end or the other.

The following observations apply for both the individual normal PDF representation cases and for the convolved PDF representation cases. The trends of method relative performance at n=2 samples (Figures 4 and 5) also generally apply at n=8 and n=32 samples (Figures 7 and 8 respectively) as will be quantified by more detailed processing in the next subsection. Nonetheless, the performance distinctions between the various methods diminish as more data samples are taken.

By observing the relative areas of the histograms it is apparent that the Pradlwarter-Schuëller (Pr-Sch) method and the Tolerance Interval method contain much greater proportions of positive results than do the Normal-Fitting and Non-Parametric methods. The detailed results processing in the next subsection will

also show that the Tolerance Interval method contains significantly more positive results than does the Pr-Sch method. In a correlated result, the Tolerance Interval method contains significantly less negative and mixed results than the other three methods. It is also immediately apparent from the +/+ histograms that at n=2 samples the Tolerance Interval method is considerably more overconservative in estimation of percentile ranges than the other methods. Pr-Sch is the next most conservative method, then the Normal-Fit method, and the least conservative is the Non-Parametric method. For larger numbers of samples the Tolerance Interval overestimation is not so pronounced relative to the other methods. In fact, at n=32 samples the Pr-Sch method has maximum overstimation errors that are slightly greater than those of the Tolerance Interval method for the normal PDF and convolution cases, as does the Non-Parametric method for the normal PDF case.

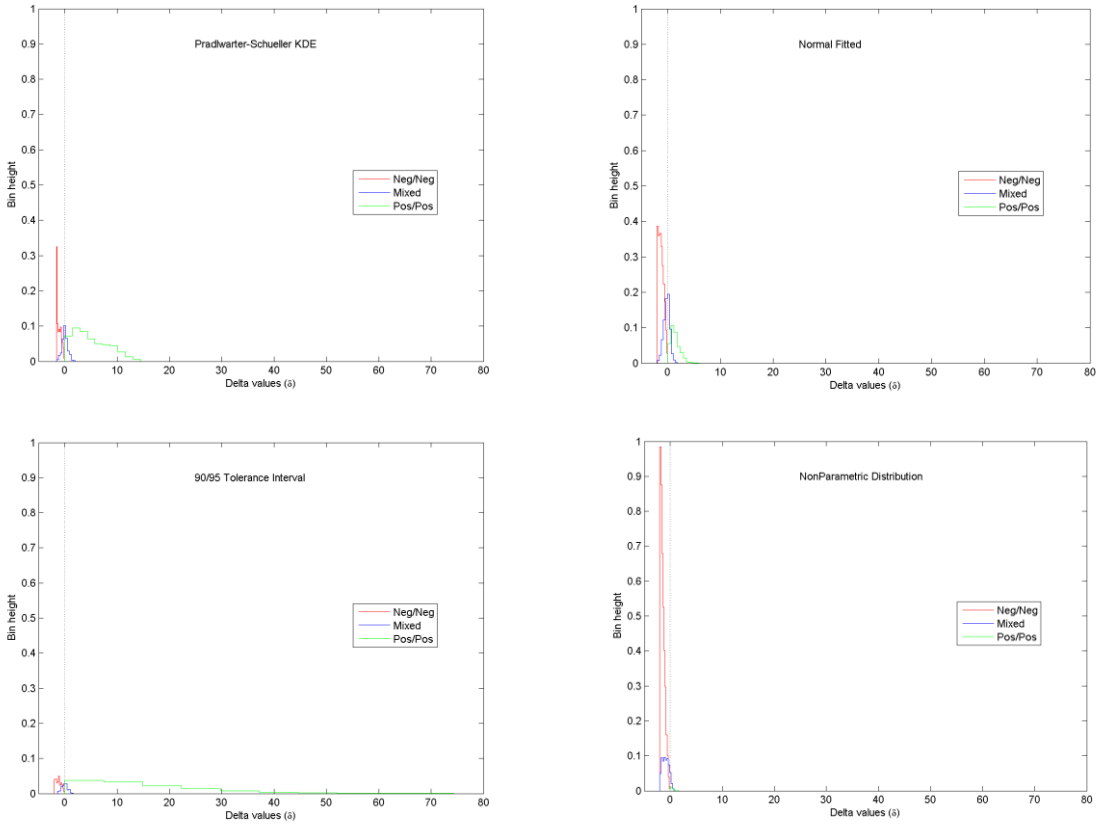

Figure 4. Histogram comparison of performance of the four estimation methods—results for n=2 samples of a normal PDF.

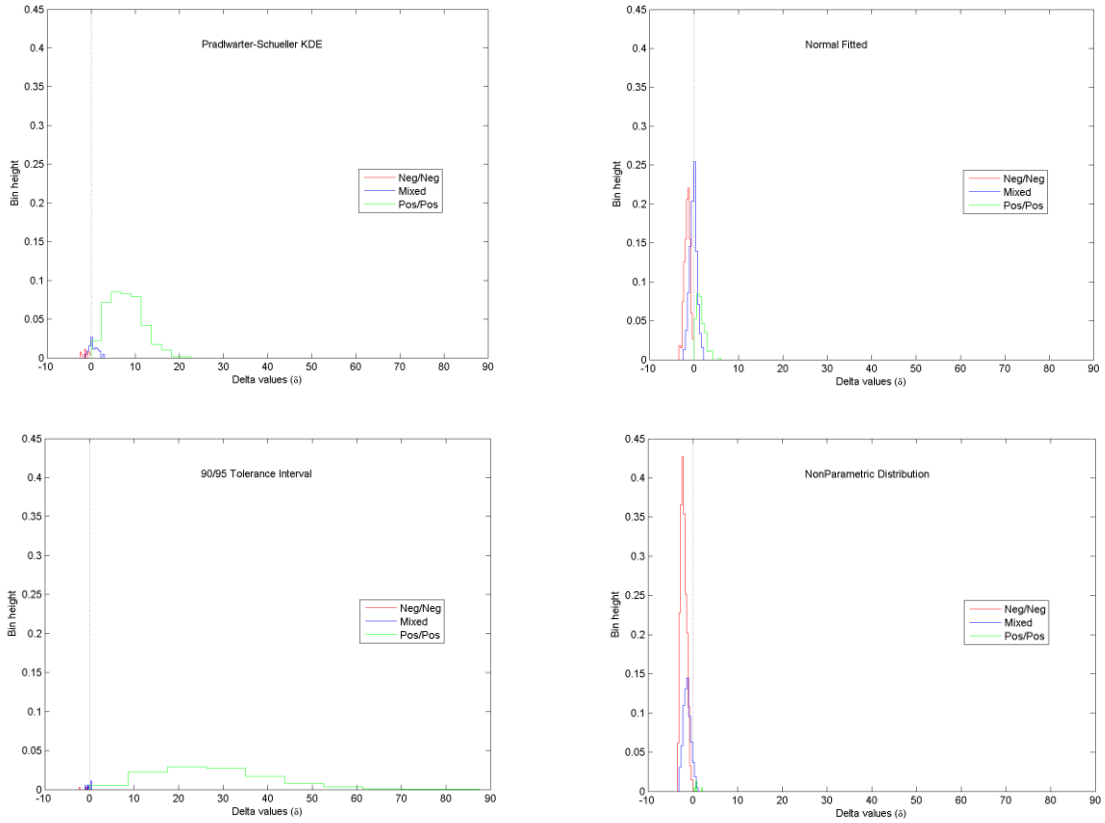

Figure 5. Histogram comparison of performance of the four estimation methods—convolution problem results for *n*=2 samples of each of the three contributing normal PDFs.

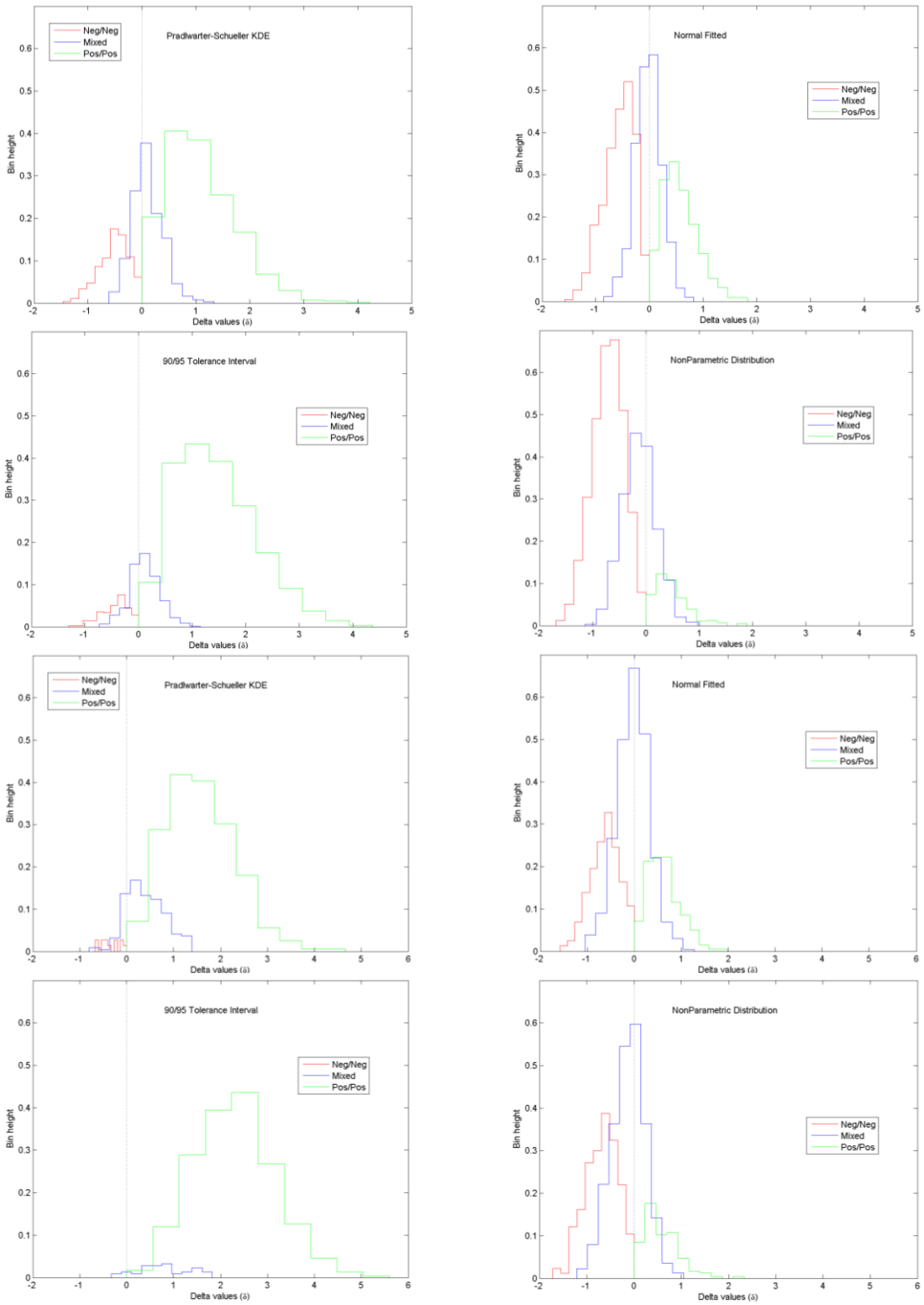

Figure 6. Histogram comparison of performance of the four estimation methods for *n*=8 samples. Top four plots are for fitting a normal PDF, bottom four plots are for 3PDF convolution problem.

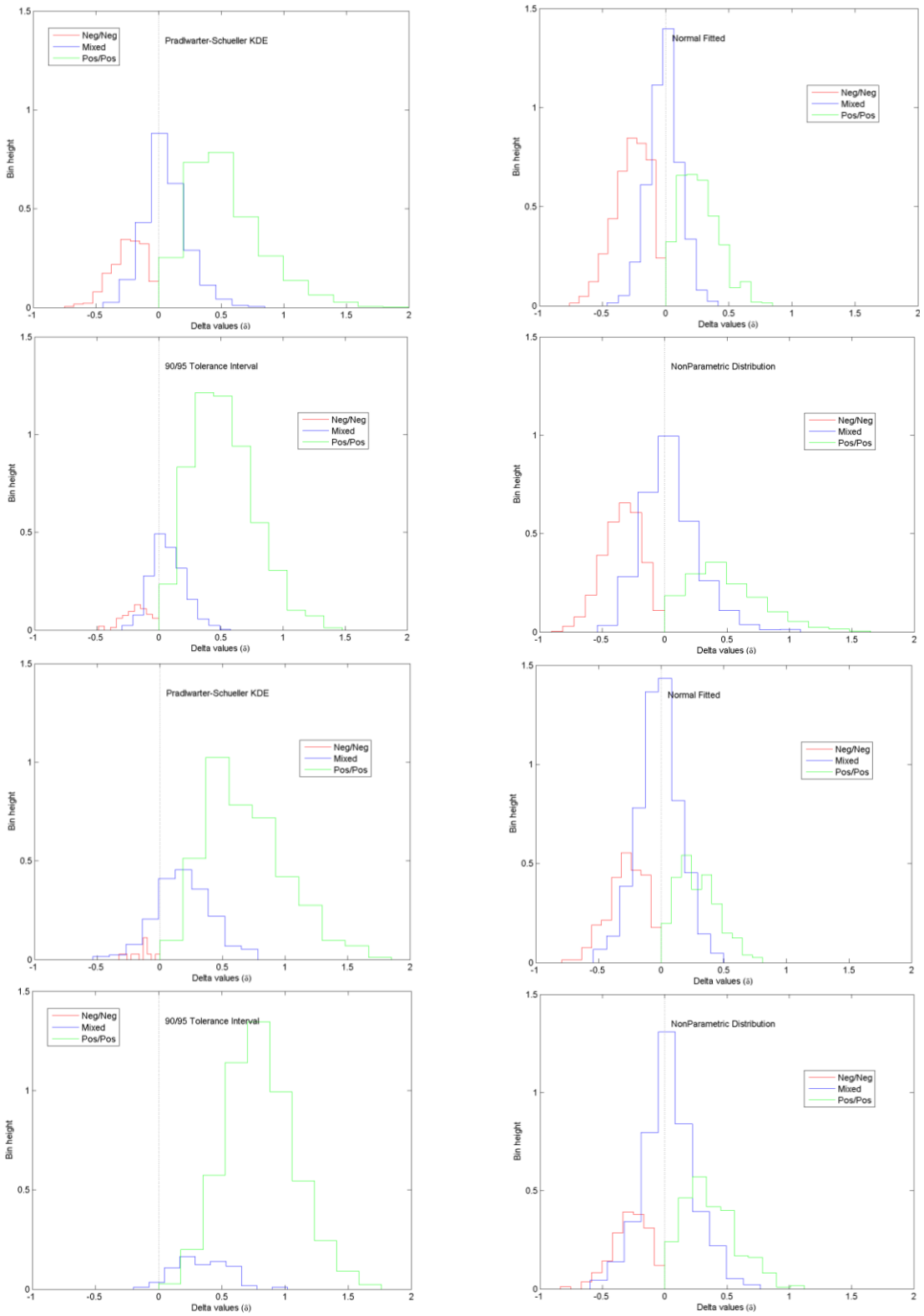

Figure 7. Histogram comparison of performance of the four estimation methods for *n*=32 samples. Top four plots are for fitting a normal PDF, bottom four plots are for 3PDF convolution problem.

B. Detailed Processing of Estimation Results

Figures 8 and 9 plot the proportion of results in +/+, -/-, and mixed categories as a function of the number of data samples used. The solid lines show the net of taking the number of +/+ results (desirable occurrences) and subtracting the number of -/- results (undesirable occurrences) to get a net number of desirable results (as a percent of the total number of trials). The higher a solid line lies on the plots, the better the method's performance with respect to this performance measure. The Tolerance Interval method always ranks best by a considerable margin according to this measure, followed by the Pr-Sch method, then generally the Normal Fitting method followed by the Non-Parametric method. The Normal-Fitting and Non-Parametric methods have more undesirable -/- results than desirable +/+ results, giving them net undesirable scores for this performance measure, even at the relatively large number of 32 samples.

The dashed lines show the percentage of trials in which mixed results occurred. Mixed results are considered to be undesirable because one end of the estimated interval does not bound the true percentile range. (Nonetheless, mixed results are not as undesirable as -/- results.) Because mixed results are undesirable, the lower the dashed line lies on the plots (i.e., the closer to zero), the better the method's performance with respect to this performance attribute. The Tolerance Interval method always performs best by a substantial margin according to this measure, followed by the Pr-Sch method, then relatively far behind are the Normal Fitting method and the Non-Parametric method (which generally performs worst).

For all methods the solid lines in Figures 8 and 9 show that as n increases from 2 to 8 to 32 samples the number of desirable +/+ results increases relative to the number of undersirable -/- results. However, for all the methods the number of less desirable mixed results also increases as the number of samples increases—somewhat muting the overall improvement with increasing number of samples.

Note that two or more methods' solid-line results could occupy the same ordinate value at a given abscissa value of say n=2 samples, yet the methods' dashed lines can have different ordinate values there. For example, let one method yield (60% +/+, 30% -/-, 10% mixed) results and another method yield (55% +/+, 25% -/-, 20% mixed) results. Then both methods have the same solid-line plotted value of 30%, reflecting an equal score of "goodness" with regard to the desirable attribute of the number of very desirable +/+ occurrences minus the number of very undesirable -/- occurrences. This tie score can be broken by considering the number of undesirable results comprising the mixed population. The method with 10% mixed results would be considered better than the method with 20% mixed results, hence the method that yielded (60% +/+, 30% -/-, 10% mixed) results is considered the overall better performer.

Accordingly, the virtual tie between the solid-line results of the Non-Parametric and Normal Fit methods at n=32 samples in Figure 8 is broken by considering their dashed-curve values at that abscissa value. The Normal Fit method has about 10% less mixed results at n=32 than does the Non-Parametric method, making the Normal Fit method's performance better overall in terms of numbers of results in the +/+, -/-, and mixed categories at n=32.

If one gives the same undesirability weight to the mixed results as is given to the -/- results, then the graph and the ranking of method performance could be made substantially simpler by just quantifying the net value [(% of +/+ results) minus (% of -/- results) minus (% of mixed results)]. The larger the value, the better the method has performed. However, it is not deemed reasonable to hold that a mixed result is as bad or undesirable as a -/- result. Therefore the two types of results should not be given the same undesirability weight in a method performance evaluation. Certainly, an appropriate weighting scheme could be devised and implemented, but in this initial paper we use a very simple zeroth-order performance ranking scheme as described in Section V.

By comparing Figs. 8 and 9 the general tendency is that the aggregation (convolution) of multiple sources of variability somewhat washes out or mitigates the relative number of undesirable -/- and mixed errors of the individual contributing sources.

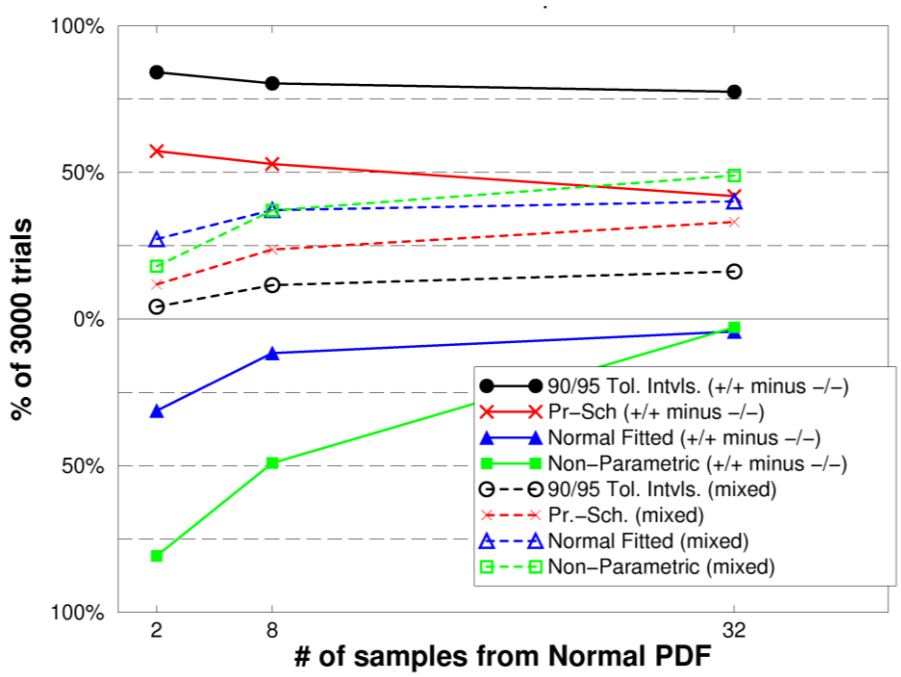

Figure 8. Comparison of proportion of results in +/+, -/-, and mixed categories (for the four estimation methods) as a function of number of data samples from a normal PDF.

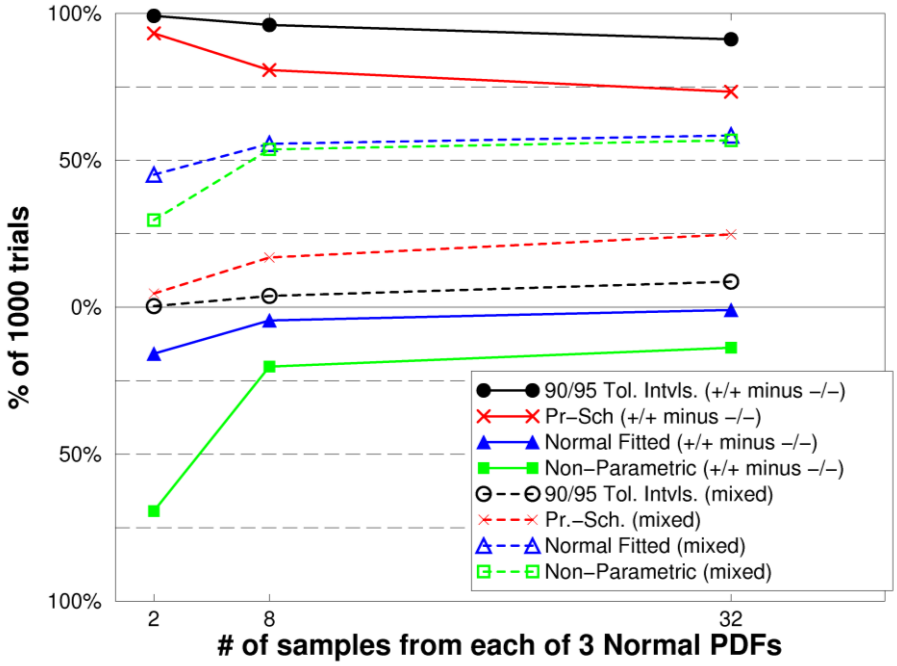

Figure 9. Comparison of number of results in +/+, -/-, and mixed categories (for the four estimation methods) as a function of number of data samples—performance in convolution problem.

Now the desirability of the methods is considered with respect to the magnitude of the overshoot and shortfall errors about the true percentile range of the target PDF. For this performance category, mixed errors are quantified by adding the absolute values of $\Delta_{U\text{-}i}$ and $\Delta_{L\text{-}i}$. This prevents large overshoot errors (positive Δ) and large shortfall errors (negative Δ) from summing to a small value $S_i = \Delta_{U\text{-}i} + \Delta_{L\text{-}i}$ and thereby projecting the appearance that the mixed errors are both small.

Figures 10 and 11 plot, for the individual normal PDF and convolved PDF representation problems, the mean predicted percentile ranges and error magnitudes for the various methods and n=2 samples. The magnitudes of the mean -/- and mixed errors are relatively similar for all methods. But for the +/+ overshoot errors, the Tolerance Interval method has large average overshoot in its large +/+ population of results. The mean overshoot error is much larger than for the Pr-Sch method, which also has a relatively large +/+ population of desirable results. The +/+ overshoot errors of the Normal-Fit and Non-Parametric methods are relatively less extreme than for the other two methods. Unfortunately the proportions of desirable +/+ results are relatively small for the Normal-Fit and Non-Parametric methods.

The relative trends in Figures 10 and 11 are generally maintained at n=8 and n=32 samples, as revealed by Figures 12 - 15. Nonetheless, the performance distinctions between the various methods diminish as more data samples are taken.

Any mitigation or washing-out of individual representation error magnitudes in the aggregation (convolution) of multiple sources of variability can be assessed by comparing Figures 12 and 13 respectively with Figures 14 and 15. Normalized error magnitudes (as a percent of the true percentile ranges) for the convolved results are generally slightly less than the individual PDF representation errors. This is generally the case for all methods, numbers of samples, and categories of +/+, -/-, and mixed overshoot and shortfall errors. However, at n=2 samples the Tolerance Interval method does not show improvement in normalized +/+ error magnitude. All in all, any effective error reductions that come with aggregation are only slight, not substantial.

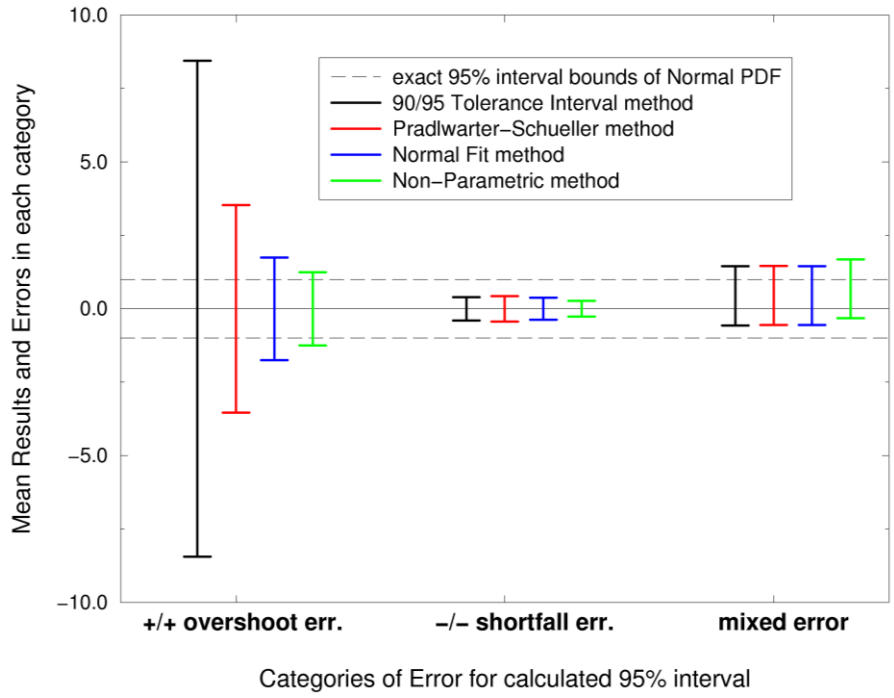

Figure 10. Mean predicted percentile ranges and +/+ overshoot error magnitude in the trials, mean -/- shortfall error magnitude in the trials, and mean mixed absolute errors for the various methods and *n*=2 samples—performance in representing a normal PDF.

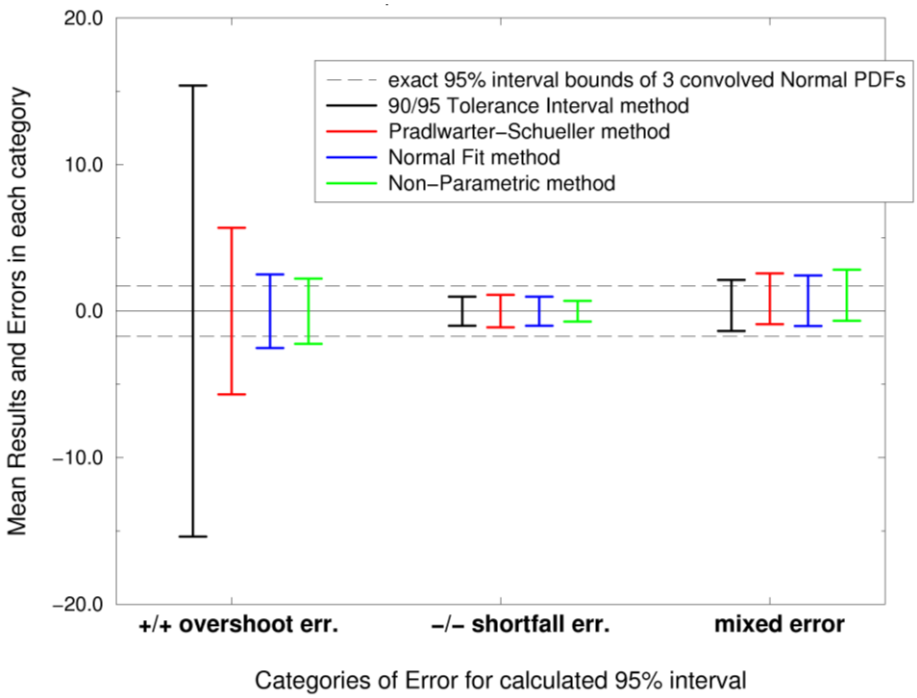

Figure 11. Mean predicted percentile ranges and +/+ overshoot error magnitude in the trials, mean -/- shortfall error magnitude in the trials, and mean mixed absolute errors for the various methods and *n*=2 samples—performance in convolution problem.

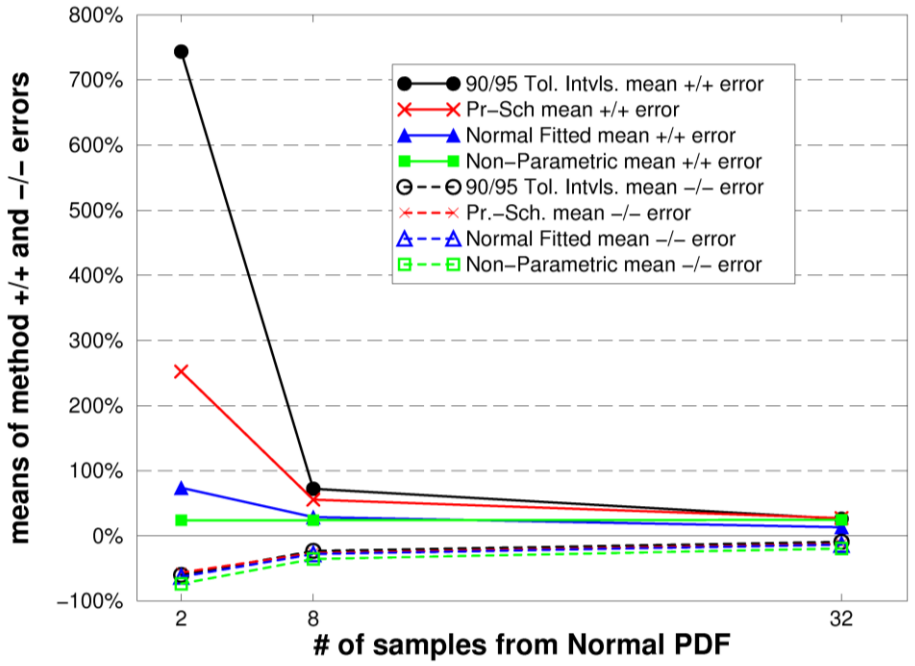

Figure 12. Mean +/+ overshoot error magnitudes and mean -/- shortfall error magnitudes, as a function of number of data samples—performance in representing a normal PDF.

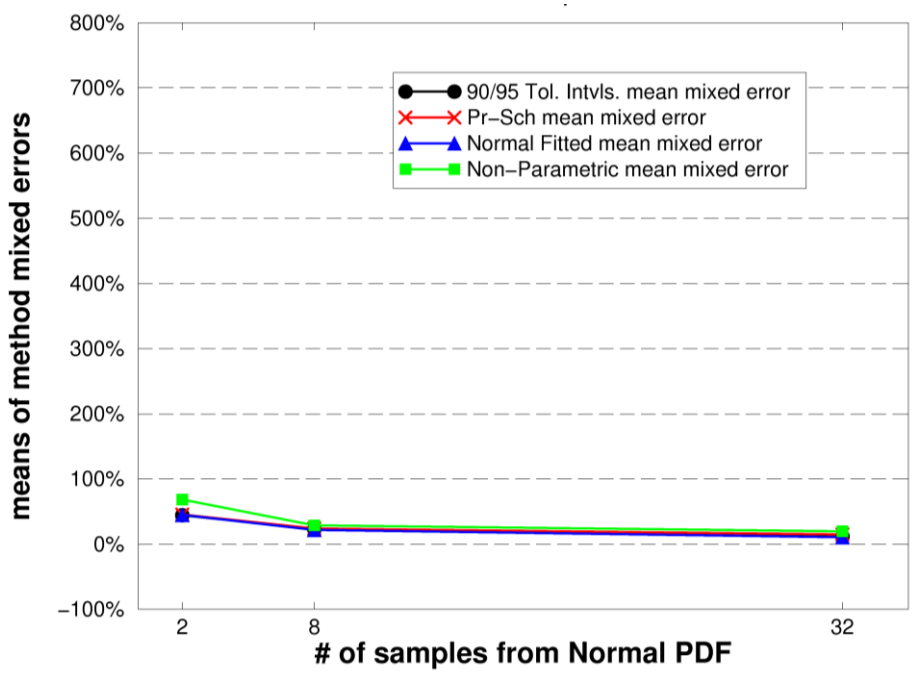

Figure 13. Mean mixed absolute errors for the various methods, as a function of number of data samples—performance in representing a normal PDF.

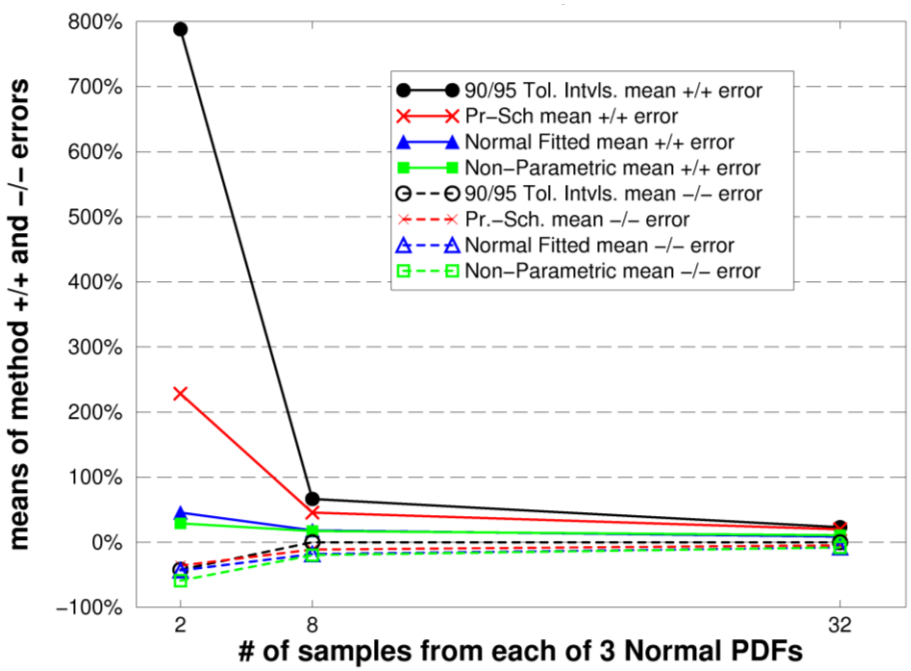

Figure 14. Mean +/+ overshoot error magnitudes and mean -/- shortfall error magnitudes, as a function of number of data samples—performance in convolution problem.

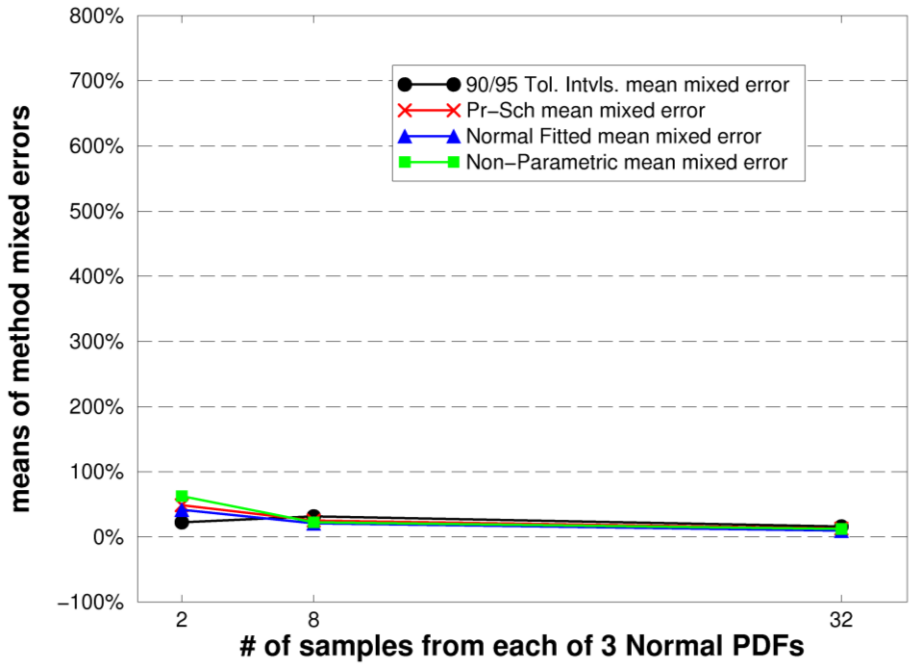

Figure 15. Mean mixed absolute errors for the various methods, as a function of number of data samples—performance in convolution problem.

V. Discussion and Conclusions

Figure 16 summarizes the relative performance of the methods according to the performance attributes discussed above and listed along the bottom of the figure. A performance rank of 1 is best and a rank of 4 is worst. A simple overall score for each method can be obtained by adding the method's rank values across the six attributes, and then dividing by six. The lowest average score identifies the best method, and the highest average score identifies the worst performing method. The average performance scores from Figure 16 are, in order of best to worst: 1.67 for the Tolerance Interval method, 2.17 for the Pr-Sch method, 2.8 for the Normal Fitting method, and 3.33 for the Non-Parametric method.

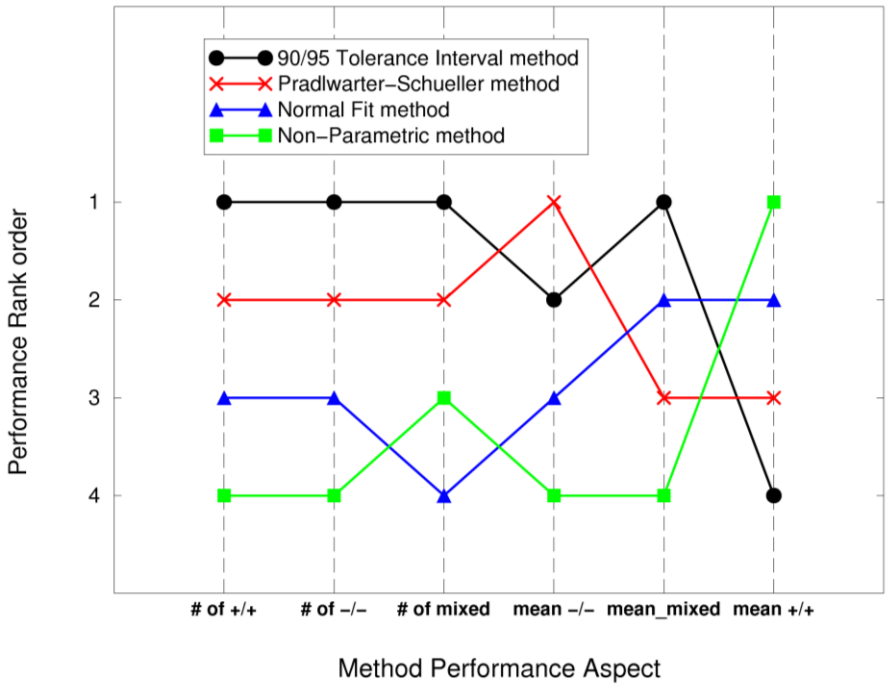

Figure 16. Performance rankings of the various methods according to the performance attributes listed at bottom of plot.

The plot and the average ranking scores were constructed for performance in representing the 95% included probability of an individual normal PDF, and for n=2 samples of data. However, the plot and ranking scores are the same for performance in the convolution problem at n=2 samples. The plot and scores are also the same for the normal-representation and convolution problems at n=8 samples. For n=32 samples, slight differences arise in the plot and method rankings for some attributes, but the method ranking order by overall average score does not come close to changing.

When the methods are judged by maximum +/+, -/-, and mixed error magnitudes instead of their mean error magnitudes considered in Figures 10-16, the rankings remain as in Figure 16, across the normal-representation and convolution problems and all sampling levels n= 2, 8, and 32. Thus, for the very simple ranking system described, the said ordering of method performance is very robust.

By the simple ranking scheme employed, the classical Tolerance Interval method performed best on the limited set of test problems tried. Furthermore, the Tolerance Interval method is simple to implement and use. However, for very few samples, the large magnitude of its overshoot errors relative to the next-ranked method of Prandlwarter & Schuëller gives reason for pause (see Figures 10, 11, 12, and 14). For moderate (8) and large (32) numbers of samples the overshoot errors of the Tolerance Interval method compared to the Pr-Sch method are not markedly larger (see Figures 12 and 14).

The very simple zeroth-order performance ranking scheme employed in this paper gives all performance aspects equal weight, which is certainly over-simplifying things. Beyond the simple consideration here of whether one method is better than another in a performance attribute, a more refined judgment of performance considers *how much better* one method does than another, say in terms of the number of realizations in the desirable +/+ category versus the very undesirable -/- category, and in terms of the magnitudes of overshoot and shortfall errors. Moreover, the various attributes can be differentially weighted according to importance to various types of analysis results and purposes. Finally, complexity of method implementation should also be factored in. Then a variety of decision theory methods could be used to help make a more rigorous decision according to user preferences, including formal utility theory which would rank the utility of different sets of attribute values, lexicographic ordering, cobweb plots, etc.

Hence, strong conclusions based on the limited study, analysis, and performance ranking scheme presented here would be premature. A more comprehensive study presently underway will give a much broader view of the relative performance of the various methods. However, one finding from this paper that is not likely to change is that the common practice of fitting data with a normal PDF is not very reliable, even if the underlying random process being sampled is Normal. For example, for n=2 samples the Normal Fitting method produced 52% very undesirable -/- errors, with a mean magnitude of 63% shortfall from the true 95-percentile range of the normal PDF. For n=8 samples, these numbers fall to 37.2% -/- errors with a mean magnitude of 28% error—still very substantial error even for the relatively large number of 8 data samples. This significant representation error is only slight mitigated in uncertainty aggregation (convolution) with multiple similarly sized and represented sources of Normal uncertainty (compare Figures 12 and 13 with 14 and 15).

Therefore it is important to instill into common engineering practice better methods for representing stochastic uncertainty when only relatively few data samples exist. The study here provides clear motivation for the use of appropriate methods for representing the combined aleatory and epistemic uncertainty associated with limited data samples of a stochastic quantity or system. Moreover, it provides an initial glimpse into the treatment options available, their implementation, and their relative performance tendencies for sparse samples of Normal random-variable data.

Acknowledgments

The authors thank the Advanced Simulation and Computing (ASC) Program of the National Nuclear Security Administration (NNSA) at Sandia National Laboratories for funding this research.

References

[1] Krishnamurthy, T., and Romero, V.J., "Construction of Response Surface with Higher Order Continuity and its Application to Reliability Engineering," paper AIAA-2002-1466, 43rd Structures, Structural Dynamics, and Materials Conference, Denver, CO, April 22-25, 2002.

[2] Romero V.J., Burkardt, J.S., Gunzburger M.D., Peterson J.S., and Krishnamurthy T., "Initial application and evaluation of a promising new sampling method for response surface generation: Centroidal Voronoi Tessellation," paper AIAA-2003-2008, 44th Structures, Structural Dynamics, and Materials Conference, Apr. 7-10, 2003, Norfolk, VA.

[3] Romero, V.J., Painton-Swiler, L., and Giunta, A.A., "Construction of response surfaces based on progressive-lattice-sampling experimental designs with application to uncertainty propagation,'' *Structural Safety* 26 (2004) 201-219.

[4] Romero V.J., Slepoy R., Swiler L.P., Giunta A.A., Krishnamurthy T., "Error estimation approaches for progressive response surfaces -more results," paper #114 in Proceedings of the 2006 International Modal Analysis Conference (IMAC XXIV), Jan. 30 - Feb.2, 2006, St. Louis, MO.

[5] Kline, S.J., and McClintock, F.A., "Describing Uncertainties in Single-Sample Experiments," *Mechanical Engineering*, Jan. 1953, pp. 3 – 8.

[6] International Organization for Standardization, *Guide to the Expression of Uncertainty in Measurement*, 1995 corrected and reprinted edition.

[7] U.S. Dept. of Commerce National Institute of Standards and Technology (NIST) Technical Note 1297, *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*, 1994 Edition, B.N. Taylor and C.E. Kuyatt.

[8] American Society of Mechanical Engineers, ASME PTC 19.1-2005, *Test Uncertainty*, 2006.

[9] Coleman, H.W., and Steele, Jr., W.G., Experimentation and Uncertainty Analysis for Engineers ─ 2nd Edition, John Wiley & Sons, New York, NY, 1999.

[10] American Society of Mechanical Engineers, V&V 20 – 2009 *Standard for Verification and Validation in Computational Fluid Dynamics and Heat Transfer*.

[11] Eldred, M.S., and J. Burkardt, "Comparison of non-intrusive polynomial chaos and stochastic collocation methods for uncertainty quantification," 47th AIAA Aerospace Sciences Meeting and Exhibit, paper AIAA-2009-0976, January 5-8, 2009, Orlando, FL.

[12] Hahn, G.J., and W.Q. Meeker, *Statistical Intervals—A Guide for Practitioners*, Wiley & Sons, 1991.

[13] Pradlwarter, H.J., and G.I. Schuëller, "The use of kernel densities and confidence intervals to copy with insufficient data in validation experiments," *Computer Methods in Applied Mechanics and Engineering*. Vol. 197, Issues 29-32, May 2008, pp. 2550-2560.

[14] Sankararaman, S., and S. Mahadevan, "Likelihood-based representation of epistemic uncertainty due to sparse point data and/or interval data," accepted for *Reliability Engineering and System Safety*, doi:10.1016/j.ress.2011.02.003.

[15] Marhadi, K., S. Venkataraman, S. Pai, "Quantifying Uncertainty in Statistical Distribution of Small Sample Data Using Bayesian Inference of Unbounded Johnson Distribution," paper AIAA-2008-1810, 10th AIAA Non-Deterministic Approaches Conference, April 7–10, 2008, Schaumburg, Illinois.

[16] Zaman, K., M. McDonald, S. Rangavajhala, S. Mahadevan, "Representation and Propagation of both Probabilistic and Interval Uncertainty," paper AIAA-2010-2853, 12th AIAA Non-Deterministic Approaches Conference, April 12–15, 2010, Orlando, Florida

[17] Johnson, N.L., "Systems of frequency curves generated by methods of translation", *Biometrika*, 1949, 36:149-176.

[18] Rosenblatt, M., "Remarks on Some Nonparametric Estimates of a Density Function," *The Annals of Mathematical Statistics*, Vol. 27, 1956, pp. 832-835.

[19] Parzen, E., "On Estimation of a Probability Density Function and More," *The Annals of Mathematical Statistics*, Vol. 33, No. 3, September 1962, pp. 1065-1076.

[20] Silverman, B.W, *Density Estimation*, Chapman and Hall, 1986.

[21] Jones, M.C., Marron, J.S., Sheather, S.J., "A Brief Survey of Bandwidth Selection for Density Estimation," *Journal of the American Statistical Association*, Vol. 91, No. 433 (March 1996), pp. 401-407.

[22] Wand, M.P. and Jones, M.C., *Kernel Smoothing*, Chapman and Hall, 1994.

[23] Alhberg, J.H., E.N. Nilson, J.L. Walsh, *The Theory of Splines and Their Applications*, Academic Press Inc., New York, NY, 1967.